Proceedings of SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.TJ22275

Injecting Visual Prototypes into CLIP-driven Long-Tailed Skin Lesion Classification

Yunzhe Pan

School of Henan Polytechnic University, Jiaozuo, China
18839681172@163.com

Abstract: Skin lesion classification plays a vital role for the early and accurate diagnosis of dermatology. Traditional methods rely on subjective clinical expertise but remain challenged by long-tailed data distributions and insufficient utilization of visual information. With the advancement of artificial intelligence, an increasing number of computer-aided approaches have been attempted for skin lesion classification tasks. In this paper, we propose DermSense, a novel framework to enhance skin lesion classification by integrating visual prototypes, LLM-generated fine-grained attributes with the CLIP model. Leveraging a dual-matching strategy and a cross-modal fusion module enriches visual textual feature representations without the need for fine-tuning the CLIP backbone network, and to significantly reduce computational overheads. Additionally, we employed a Logits-Adjusted loss function to address the class imbalance. Extensive experiments on ISIC2018 and Derm7pt datasets demonstrate that DermSense significantly outperforms state-of-the-art methods in balanced accuracy, particularly for melanoma classification. Our results highlight the potential of DermSense to advance skin disease diagnosis while maintaining scalability and efficiency.

Keywords: Skin Lesion Classification, Vision-Language Models, Long-Tailed Learning

1. Introduction

Skin lesions span diverse dermatological conditions including benign growths and malignant neoplasms such as melanoma, basal cell carcinoma, and squamous cell carcinoma [1]. These pathologies present global healthcare challenges due to their epidemiological prevalence and risks of severe outcomes from diagnostic delays. Traditional approaches dependent on clinical expertise remain constrained by human perceptual limitations, potentially causing diagnostic inconsistencies [2]. Recent advancements in artificial intelligence, particularly deep learning architectures [3-6], have revolutionized dermatological image analysis through automated feature extraction, significantly improving diagnostic precision. The medical adaptation of Contrastive Language–Image Pretraining (CLIP) models [7-9] leveraging its powerful image-text matching capabilities to enhance downstream applications. These adaptations have demonstrated significant improvements in performance by effectively integrating multi modal information across diverse medical imaging challenges. Specifically, in skin lesion classification, several studies have utilized CLIP to incorporate semantic concepts of different skin diseases [10]. This paradigm enhances discriminative capabilities for lesion characterization while maintaining computational efficiency, highlighting its transformative potential for clinical dermatology.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.TJ22275

Challenges persist in CLIP-based skin lesion classification. First, CLIP relies highly on image-text matching for downstream zero shot prediction. However, the class name and corresponding concepts of a skin lesion are obscure. Subsequently, conceptual overlaps and underutilized visual patterns remain weak points, especially when encountering long-tailed data distribution. Additionally, task-specific fine-tuning requirements incur prohibitive computational costs, limiting clinical deployment. Consequently, there is a need for more efficient strategies that leverage CLIP's capabilities without compromising accuracy or scalability.

In light of the above concerns, we propose a novel frame work, DermSense, for skin lesion diagnosis. DermSense leverages the commonsense knowledge of Large Language Models (LLMs) to generate highly discriminative visual attributes for various skin diseases, thereby enhancing the categorical information used in classification. To address the challenge of insufficient visual information utilization, we introduce visual prototypes for each disease category, which serve to complement the image-text matching process with robust image-to-image matching. This dual-matching approach ensures a more nuanced and accurate differentiation between visually similar skin lesions. Additionally, we utilize a cross-modality fusion module based on the attention mechanism to enrich the feature representations without the need to fine-tune the CLIP backbone, thereby maintaining computational efficiency. To further tackle the issue of imbalanced data distributions commonly found in skin lesion datasets, we implement a Logits-Adjusted loss function [11]. This loss function effectively mitigates the long-tail problem by calibrating the predicted logits with the prior probability of each category. This mechanism ensures balanced learning across categories without distorting the original data distribution, effectively reconciling sensitivity and specificity in imbalanced dermatological scenarios. By combining these strategies, DermSense enhances both the accuracy and scalability of skin lesion classification, providing a robust framework for improved diagnostic performance.

Our main contributions are summarized as follows:

- We design a novel framework called DermSense for more reliable and efficient skin lesion classification, preserving the text-image alignment strengths of the CLIP model but also mitigating the issue of visual insufficiency.
- We propose to incorporate the distinguished attributes generated by LLM with class names. We also create visual prototypes for each skin disease to supplement the insufficient usage of visual information. To enrich representations of different modalities, we propose to perform cross-modal attention on the feature space.
- Experiments on two commonly used skin lesion datasets demonstrate that our method achieve a superior classification performance for melanoma while ensuring computational efficiency.

2. Related Work

2.1. Skin Lesion Classification

Classification of skin lesions by machine learning has long been a trend. Automatic classification of skin lesions is urgently needed in clinical examination, which is aimed to assist physicians efficiently and provide rapid diagnoses [1]. While suffering from access to abundant datasets, both traditional machine learning and deep learning methods are adequately developed. Traditional machine learning methods, such as multilayer perceptron (MLP), support vector machines (SVM), and random forests, typically follow workflows involving preprocessing, segmentation, feature extraction, and classification. For instance, F-MLP [2] employs fuzzy logic to optimize lesion border prediction, while Melbin et al. [3] integrate modified asymmetry, border, color, and diameter (ABCD) features with SVM for improved accuracy. Deep learning approaches predominantly rely on convolutional

DOI: 10.54254/2755-2721/2025.TJ22275

neural networks (CNNs). Kawahara et al. [12] proposed a CNN-based system for multi-resolution classification, and Zanddizari et al. [5] enhanced CNN performance through region-of-interest preprocessing. However, limited training data poses risks of overfitting [6], and scalability remains uncertain for larger datasets.

2.2. Vision-Language Models in Biomedical Domain

Vision-language models (VLMs), such as CLIP and its biomedical variants BiomedCLIP [7] and PubMedCLIP [8], learn joint image-text representations to bridge domain gaps in medical tasks. These models demonstrate strong zero-shot capabilities, as seen in CheXzero [14] for chest X-ray classification, Clip-lung [15] for malignancy prediction in lung CT scans, and CoOpLVT [16] for diabetic retinopathy detection using retinal fundus images. CLIP variants like GLIP [17] excel in lesion detection across modalities, while VLPMNuD [18] achieves zero-shot nuclei segmentation in histopathology images and MCGN [19] leverages CLIP's vision encoder to enhance chest X-ray report generation accuracy. Despite their versatility, VLMs underperform task-specific models and lack interpretability. Our method addresses these limitations by integrating discriminative attributes into CLIP and employing a cross-modal fusion module, achieving superior melanoma classification accuracy with low computational overhead.

3. Method

As illustrated in Figure 1, our method begins with the preprocessing of both image and text modalities, followed by the pre-computation of visual prototypes and textual features corresponding to each class. These representations are then enriched through an interaction mechanism utilizing multi-head self-attention, enabling comprehensive feature-prototype fusion. To address the long-tail distribution challenge, the prediction weights are further refined using a logit-adjusted loss, ensuring balanced and robust classification performance, mitigating the dominance of head classes during training and improving robustness for rare categories.

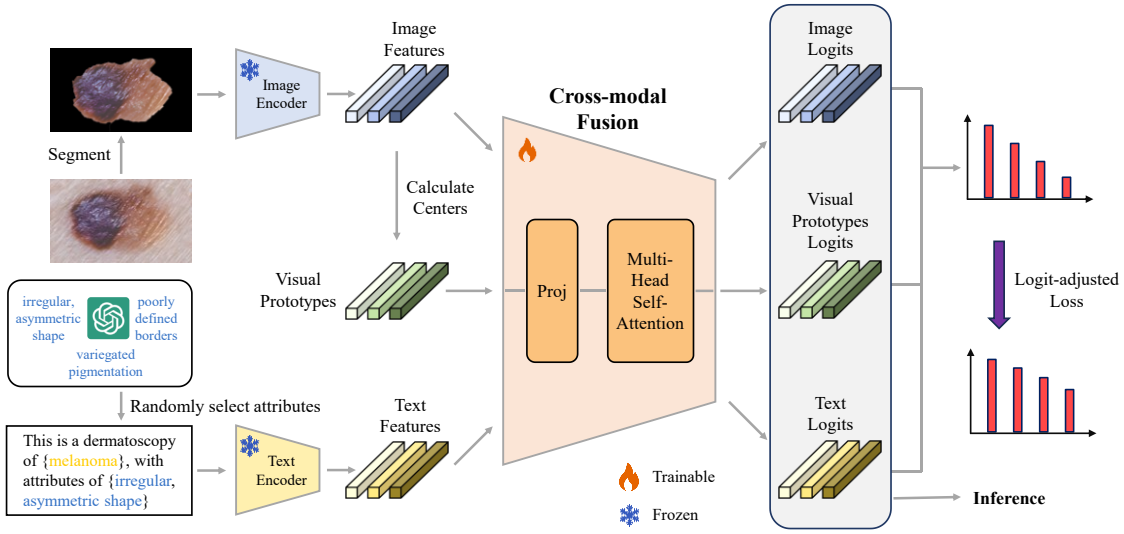

Figure 1: Overview of DermSense. The pipeline begins with pre-processing the image and text modalities, including segmenting skin lesion images using an off-the-shelf segmentation model and generating distinctive attributes. Next, visual prototypes and textual features for each class are pre-computed and stored. These features and prototypes are then integrated through cross-modal attention, yielding prediction logits. To further address the long-tail distribution challenge, a logit-adjusted loss is applied.

DOI: 10.54254/2755-2721/2025.TJ22275

3.1. Revisit CLIP

Contrastive Language-Image Pretraining (CLIP) is a groundbreaking vision-language model that forms the corner stone of our approach. It consists of an image encoder, $f_I(x)$, and a text encoder, $f_T(t)$. The image encoder can utilize architectures such as ResNet [20] or ViT [21], while the text encoder is based on the Transformer architecture [22].

Trained on a large-scale dataset of 400 million image-text pairs using a contrastive loss, CLIP learns to align semantically corresponding pairs and separate mismatched ones within a shared embedding space. For zero-shot image classification, given an image x and a set of class descriptions $\{t_1, \ldots, t_K\}$, CLIP extracts image features $X = f_I(x)$ and textual features $T = \{T_1, \ldots, T_K\} = \{f_T(t_1), \ldots, f_T(t_K)\}$. The predicted class for x is determined as:

$$y_{prep} = arg \max_{i \in [K]} cos(X, T_i), \tag{1}$$

reframing traditional image classification as an image-text matching problem.

Despite its strengths, its direct application to skin lesion analysis remains constrained by two inherent limitations. In datasets with long-tailed distributions, severe class imbalance—where rare lesion categories are vastly underrepresented—leads to biased feature learning and degraded generalization for minority classes. Furthermore, the model's reliance on generic textual descriptions often inadequately encodes the nuanced visual characteristics of certain lesions, resulting in suboptimal visual-semantic alignment. These challenges motivate the design of our method, which seeks to address these shortcomings effectively.

3.2. Attributes generated by LLM

To augment discriminability in skin lesion classfication, our approach employs GPT-4o [23] to synthesize concise yet semantically distinctive attributes for each lesion category. This strategy addresses the challenge of distinguishing visually analogous samples by explicitly encoding subtle pathological differences that conventional visual analysis might neglect. To further enhance the distinction between lesions, we enforce a uniqueness constraint during attribute generation, ensuring that GPT-4o produces mutually orthogonal descriptors to mitigate inter-class concept overlap. This mitigates the risk of concept overlap, which could lead to confusion when classifying skin lesions that share similar visual features. Examples of the generated attributes are illustrated in Figure 2.

3.3. Feature-level Cross-modal Fusion

To enrich the representation capabilities of our model, we introduce a feature-level cross-modal fusion module that employs cross-modal attention to effectively leverage both visual and textual information. To enhance the module's efficiency and practicality, we treat the model as a black box and perform optimizations directly within the feature space.

Our approach proceeds in the following steps. First, following standard practices in this domain, we use an off the-shelf segmentation model to isolate lesion areas. These segmented images are then processed by the image encoder to extract image features, denoted as X. Then, we pre-compute and store the visual prototypes, $\sum_{j=1}^{K} V_j$, and the textual features $\sum_{j=1}^{K} T_j$, for each class. The extracted image features and pre-computed prototypes are fed into their respective linear projection layers, P_I

DOI: 10.54254/2755-2721/2025.TJ22275

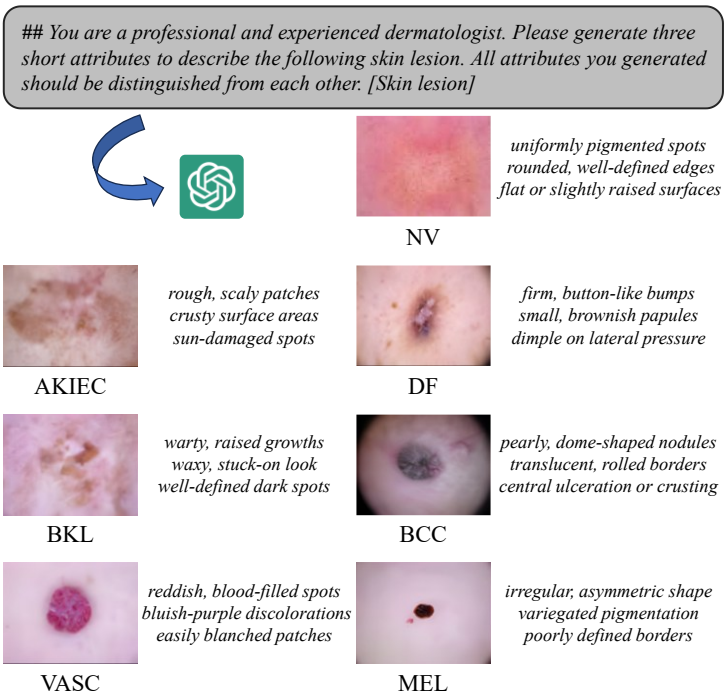

Figure 2: Examples of prompts and attributes generated by the LLMThe LLM is guided to generate attributes that are sufficiently distinct from one another, reducing the risk of concept overlap and ensuring clarity in representation.

and P_T. This step aligns the two modalities to facilitate downstream adaptation. The class prediction is then computed as:

$$z(y = i \mid X) = \frac{\exp(\cos(P_I(X), P_T(T_i))/\tau_t)}{\sum_{j=1}^{K} \exp(\cos(P_I(X), P_T(T_j))/\tau_t)} \tag{2}$$

where τ_t is the temperature for image-text matching, $cos\left(P_I(X), P_T(T_i)\right)$ denotes the cosine similarity between projection-aligned image and text embeddings in a joint space, $exp(\,)$ denotes the exponential function, amplifying cosine similarity between aligned image-text feature, Krefers the number of categories in the data set. To further enhance the richness of the features, we enable interaction between the two modalities through cross-modal attention. Specifically, we concatenate the image features, visual prototypes, and textual features, and feed them into a multi-head self-attention module [22], leveraging its ability to establish long-range dependencies within embeddings. we denote $Self - Attn(\,)$ as the self-attention operation, and the output is given by:

$$X', V', V' = \text{Self} - \text{Attn}([P_I(\text{X}), P_I(\text{V}), P_T(\text{T})]). \tag{3}$$

The self-attention operation is permutation-invariant, allowing us to slice the output to recover the enriched features and prototypes. With these enriched features, predictions are made using both the visual prototypes and textual features as follows:

$$z_v(y = i \mid X) = \frac{\exp(\cos(X', V_i')/\tau_v)}{\sum_{j=1}^{K} \exp\left(\cos\left(X', V_j'\right)/\tau_v\right)}, \tag{4}$$

$$z_T(y = i \mid X) = \frac{\exp(\cos(X', T_i')/\tau_t)}{\sum_{j=1}^{K} \exp\left(\cos\left(X', T_j'\right)/\tau_t\right)}, \tag{5}$$

DOI: 10.54254/2755-2721/2025.TJ22275

where τ_v denotes the temperature for visual modality. By ensembling these predictions, our method effectively integrates visual and textual information, leveraging their complementary strengths to enhance performance.

3.4. Overall Objective

To further address the challenges posed by long-tailed distributions in skin lesion datasets, we adopt the Logit Adjusted (LA) loss [11], a technique designed to mitigate class imbalance effectively. Let z_i denote the predicted logit for class i, the LA loss is defined as:

$$L_{la}(z_i, y = i) = -\log \frac{\exp(z_i + \log\, p(y=i))}{\sum_{j=i}^{K} \exp\,(z_j + \log\, p(y=j))} \tag{6}$$

where $p(y = i)$ denotes the class prior probability. Given n_i, the number of samples for class i, $p(y = i)$can be estimated as $\frac{n_i}{\sum_{k=1}^{K} n_k}$. This formulation allows us to effectively handle class imbalance across the base classes, ensuring fairer learning outcomes. The overall loss function is constructed by applying the LA loss to the logits calculated in Equations 2, 4, and 5 and aggregating the results. Let z_P, z_V, and z_T,denote the logits from these equations, respectively. The final loss objective L for training is expressed as:

$$L = L_{la}(z_P, y) + L_{la}(z_V, y) + L_{la}(z_T, y). \tag{7}$$

During inference, we aggregate the logits obtained after cross-modal fusion, combining the contributions from visual and textual modalities as $z = z_V + z_T$. This strategy ensures robust and balanced predictions.

4. Experiment

4.1. Experimental Settings

Table 1: Dataset statistics.

Dataset	Classes	Train Size	Validation Size	Test Size
ISIC2018[24],[25]	7	8012(890)	2003(223)	1511(171)
Derm7pt[26]	2	346(90)	161(61)	320(101)

Numbers between rounded brackets represent the number of examples of melanoma in the split.

Table 2: Evaluations on skin lesion classification performance.

Methods	Trainable Parameters	Training Time	Datasets					
			ISIC2018			Derm7pt		
			Sens.	Spec.	BAcc	Sens.	Spec.	BAcc
CLIP + Linear Classifier (ViT-L/14)	307M	-5h	24.5	92.1	58.3	36.9	91.3	64.1
MONET + Linear Classifier (ViT-L/14)	307M	-5h	35.2	**93.2**	64.2	46.8	90.8	68.8
CBI-VLM + Linear Classifier (ViT-L/14)	1.05M	-1min	43.3	91.7	67.5	**58.8**	87.6	73.2
Ours (ViT-L/14)	4.19M	-13min	52.0	87.6	69.8	55.1	93.7	74.4

DOI: 10.54254/2755-2721/2025.TJ22275

Table 2: (continued).

Ours (ViT-B/16)	2.36M	10min	55.7	88.5	**72.1**	56.7	**94.5**	**75.6**
Ours (ViT-B/32)	2.36M	-10min	**56.6**	86.2	71.4	56.8	93.6	75.2
Ours (ResNet50)	4.19M	-4min	55.1	88.1	71.6	55.9	93.9	74.9
Ours (ResNet101)	4.19M	-7min	54.7	87.9	71.3	56.3	94.1	75.2

Datasets and Evaluation Metrics. We evaluate our proposed method, DermSense, on two publicly available dermoscopic datasets: ISIC2018 [24,25]. and Derm7pt [26]. The ISIC2018 dataset consists of dermoscopic images encompassing various types of skin lesions, including melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, and vascular lesions. The Derm7pt dataset comprises both clinical and dermoscopic images. Following [10], we filter the images to include only nevus and melanoma lesions. For dataset splits, we adhere to the partitioning protocols established in [26] for Derm7pt [27] and for ISIC2018. Detailed dataset statistics, including train/validation/test splits, are presented in Table 1.

Given melanoma's high mortality rate among skin diseases, our inference focuses exclusively on determining whether a given image represents a melanoma lesion. For evaluation, we report Balanced Accuracy (BAcc) as the primary metric for handling imbalanced datasets. BAcc is calculated as the arithmetic mean of Sensitivity (True Positive Rate) and Specificity (True Negative Rate).

Baselines. We compare DermSense with three state-of-the art skin lesion classification methods: 1) CLIP [13]: Computes similarity between image visual descriptors and textual features of disease labels. 2) MONET [28]: A vision-language foundation model pre-trained on 105,550 dermatology images paired with clinical text descriptions from medical literature. 3) CBI-VLM [10]: Adapts CLIP via embedding learning and concept-driven textual embeddings for skin lesion diagnosis.

Implementation Details. For lesion segmentation, we employ DeepLabv3 [29] that fine-tuned on medical image segmentation datasets. Our models are trained for 50 epochs on each dataset using the SGD optimizer with a cosine annealing learning rate scheduler with the following hyper-parameters: a batch size of 128, a learning rate of 3e-4, a weight decay of 5e-4, and a momentum of 0.9. The temperature parameter τ_t for image-text matching is set to 0.01, following the original CLIP design. For image-image matching, the temperature parameter τ_v is tuned separately for each dataset, with values selected from the range 0.005, 0.01, 0.02, 0.05, 0.1. We report related baseline results from the CBI-VLM [10] original publication.

4.2. Skin Lesion Classification Performance

Table 2 shows the performance comparison of our proposed method with existing approaches on the ISIC2018 and Derm7pt datasets, using Sensitivity (Sens.), Specificity (Spec.), and Balanced Accuracy (BAcc) as evaluation metrics. Among the baseline methods, MONET + Linear Classifier

Table 3: Ablation study on modules we use.

Attr.	LA Loss	Vis. Proto.	CM Fusion	BAcc
				58.3
√				61.9
	√			65.7
√	√	√		70.4
√	√		√	68.5
√	√	√	√	72.1

DOI: 10.54254/2755-2721/2025.TJ22275

Table 4: Cross-dataset evaluation.

Methods	ISIC2018 → Derm7pt		
	Sens.	Spec.	BAcc
CLIP + Linear Classifier	7.2	27.4	17.3
MONET + Linear Classifier	17.0	35.8	26.4
CBI-VLM + Linear Classifier	9.3	42.5	25.9
Ours (ViT-B/16)	26.3	41.1	33.7

achieves the highest specificity on ISIC2018 (93.2%) but exhibits lower sensitivity (35.2%), while CBI-VLM + Linear Classifier improves sensitivity (43.3%) and BAcc (67.5%) with reduced training time. On Derm7pt, CBI-VLM demonstrates further gains in sensitivity (58.8%), surpassing other baselines. In contrast, our method consistently achieves superior performance with a better trade-off between sensitivity and specificity. Notably, ViT-B/32 attains the highest balanced accuracy (56.6%) on ISIC2018, while ViT-B/16 achieves the best BAcc (75.6%) and specificity (94.5%) on Derm7pt. Moreover, ResNet50 and ResNet101 perform competitively across both datasets, further validating the robustness of our approach. Importantly, our models significantly reduce training time compared to baselines, requiring only 4–13 minutes, while existing methods like CLIP and MONET demand 5 hours. The results demonstrate that DermSense not only outperforms SOTA methods in balanced accuracy but also achieves substantial training efficiency with various architectures.

4.3. Ablation Study

Table 3 presents the results of the ablation study on the individual components of our method, including attributes generated by LLM (Attr.), logits-adjusted loss (LA Loss), visual prototypes (Vis. Proto.), and cross-modal fusion (CM Fusion). Starting with only the attribute component, the model achieves a baseline balanced accuracy (BAcc) of 58.3%, where GPT-4o-generated attributes are encoded as textual prompts and aligned with CLIP's text encoder Incorporating. LA Loss improves performance to 61.9%, as the loss adjusts logits using class priors, explicitly down-weighting head classes and amplifying gradients for tail classes during training. while adding the visual prototype further enhances BAcc to 65.7%, demonstrating the effectiveness of these modules in improving the model's classification capability. The inclusion of cross-modal fusion significantly boosts performance, achieving a BAcc of 70.4% when combined with visual prototypes and 68.5% when used with LA Loss. The full combination of all components—attributes, LA Loss, visual prototypes, and cross-modal fusion—achieves the highest BAcc of 72.1%, validating the benefits of each module. These results highlight that each module contributes to the overall performance improvement, with cross-modal fusion playing a critical role in achieving optimal balanced accuracy.

4.4. Cross-dataset Evaluation

Table 4 highlights the results of cross-dataset evaluation from ISIC2018 to Derm7pt, where our method, utilizing the ViT-B/16 backbone enriched with GPT-4o-generated supplementary attributes, achieves a BAcc of 33.7%, outperforming alternatives like MONET (26.4%) and CBI-VLM (25.9%). Our approach also excels in sensitivity (26.3%) and specificity (41.1%), showcasing a stronger ability to generalize across datasets. These improvements stem from the GPT 4o-generated attributes, which provide an enriched semantic understanding of dermatological concepts. By integrating these attributes, our model better captures nuanced skin disease features, enabling it to bridge domain gaps effectively. This demonstrates the critical role of supplementary attributes in enhancing cross-dataset performance.

DOI: 10.54254/2755-2721/2025.TJ22275

4.5. Discussion

Our study, while demonstrating improved classification ac curacy in skin lesion analysis, has limitations. The clinical applicability of our framework is yet to be extensively validated, particularly with diverse datasets encompassing a broader range of skin tones and disease types. Additionally, the framework's integration into clinical practice needs further exploration to ensure it meets the requirements for transparency, explainability, and practicality in real-world scenarios.

5. Conclusion

In this paper, we present DermSense, a framework for skin lesion classification that addresses long-tailed data and visual-textual misalignment that provides a robust and scalable solution for skin lesion classification. Our method integrates visual prototypes, attributes generated by LLMs, and a cross-modal fusion module, which together significantly improve the differentiation and classification of visually similar skin lesions. Our approach achieves higher accuracy and reduced training times compared to existing methods, as demonstrated on the ISIC2018 and Derm7pt datasets. Future work will focus on expanding the diversity of the dataset and further enhancing the model's interpretability and clinical applicability.

References

[1] *R. E. Foraker, B. Kite, M. M. Kelley, A. M. Lai, C. Roth, M. A. Lopetegui, A. B. Shoben, M. Langan, N. L. Rutledge, and P. R. Payne, "Ehr-based visualization tool: adoption rates, satisfaction, and patient outcomes," eGEMs, vol. 3, no. 2, 2015.Bootorabi, F., Haapasalo, J., Smith, E., Haapasalo, H. and Parkkila, S. (2011) Carbonic Anhydrase VII—A Potential Prognostic Marker in Gliomas. Health, 3, 6-12.*

[2] *A.-R. Ali, J. Li, S. Kanwal, G. Yang, A. Hussain, and S. Jane O'Shea, "A novel fuzzy multilayer perceptron (f-mlp) for the detection of irregularity in skin lesion border using dermoscopic images," Frontiers in medicine, vol. 7, p. 297, 2020.*

[3] *K. Melbin and Y. J. V. Raj, "Integration of modified abcd features and support vector machine for skin lesion types classification," Multimedia Tools and Applications, vol. 80, no. 6, pp. 8909–8929, 2021.*

[4] *C.R. Dhivyaa, K. Sangeetha, M. Balamurugan, S. Amaran, T. Vetriselvi, and P. Johnpaul, "Skin lesion classification using decision trees and random forest algorithms," Journal of Ambient Intelligence and Humanized Computing, pp. 1–13, 2020.*

[5] *H. Zanddizari, N. Nguyen, B. Zeinali, and J. M. Chang, "A new preprocessing approach to improve the performance of cnn-based skin lesion classification," Medical & Biological Engineering & Computing, vol. 59, no. 5, pp. 1123–1131, 2021.*

[6] *Kwasigroch, M. Grochowski, and A. Mikołajczyk, "Neural architect-ture search for skin lesion classification," Ieee Access, vol. 8, pp. 9061 9071, 2020.*

[7] *S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, et al., "Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," arXiv preprint arXiv:2303.00915, 2023.*

[8] *S. Eslami, C. Meinel, and G. De Melo, "Pubmedclip: How much does clip benefit visual question answering in the medical domain?," in Findings of the Association for Computational Linguistics: EACL 2023, pp. 1181–1193, 2023.*

[9] *Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," arXiv preprint arXiv:2210.10163, 2022.*

[10] *Patrício, L. F. Teixeira, and J. C. Neves, "Towards concept-based interpretability of skin lesion diagnosis using vision-language models," in 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–5, IEEE, 2024.*

[11] *K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," arXiv preprint arXiv:2007.07314, 2020.*

[12] *J. Kawahara and G. Hamarneh, "Multi-resolution-tract cnn with hybrid pretrained and skin-lesion trained layers," in International workshop on machine learning in medical imaging, pp. 164–171, Springer, 2016.*

[13] *Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in International conference on machine learning, pp. 8748–8763, PMLR, 2021.*

DOI: 10.54254/2755-2721/2025.TJ22275

[14] Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar, "Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning," Nature Biomedical Engineering, vol. 6, no. 12, pp. 1399–1406, 2022.

[15] Y. Lei, Z. Li, Y. Shen, J. Zhang, and H. Shan, "Clip-lung: Textual knowledge-guided lung nodule malignancy prediction," in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 403–412, Springer, 2023.

[16] S. Baliah, F. A. Maani, S. Sanjeev, and M. H. Khan, "Exploring the transfer learning capabilities of clip in domain generalization for diabetic retinopathy," in International Workshop on Machine Learning in Medical Imaging, pp. 444–453, Springer, 2023.

[17] M. Guo, H. Yi, Z. Qin, H. Wang, A. Men, and Q. Lao, "Multiple prompt fusion for zero-shot lesion detection using vision-language models," in International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 283–292, Springer, 2023.

[18] Y. Wu, Y. Zhou, J. Saiyin, B. Wei, M. Lai, J. Shou, Y. Fan, and Y. Xu, "Zero-shot nuclei detection via visual-language pre-trained models," in International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 693–703, Springer, 2023.

[19] Z. Wang, M. Tang, L. Wang, X. Li, and L. Zhou, "A medical semantic assisted transformer for radiographic report generation," in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 655–664, Springer, 2022.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

[21] Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[22] Vaswani, "Attention is all you need," Advances in Neural Information Processing Systems, 2017.

[23] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[24] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 168 172, IEEE, 2018.

[25] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific data, vol. 5, no. 1, pp. 1–9, 2018.

[26] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 2, pp. 538–546, 2018.

[27] Barata, V. Rotemberg, N. C. Codella, P. Tschandl, C. Rinner, B. N. Akay, Z. Apalla, G. Argenziano, A. Halpern, A. Lallas, et al., "A reinforcement learning model for ai-based decision support in skin cancer," Nature Medicine, vol. 29, no. 8, pp. 1941–1946, 2023.

[28] Kim, S. U. Gadgil, A. J. DeGrave, J. A. Omiye, Z. R. Cai, R. Daneshjou, and S.-I. Lee, "Transparent medical image ai via an image–text foundation model grounded in medical literature," Nature Medicine, vol. 30, no. 4, pp. 1154–1165, 2024.

[29] L.-C. Chen, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017